Comparing Speed-of-Sight studies using rendered vs. natural images

Karissa Sanbonmatsu, Michael Ham, Ryan Bennett, Shawn Barr, Vadas Gintautas, Cristina Renaudo, Steven Brumby, John George, Garrett Kenyon, Luis Bettencourt

Los Alamos National Laboratory

Abstract

Viewpoint invariant object recognition is both an essential capability of biological vision and a key goal of computer vision systems. A critical parameter in biological vision is the amount of time required to recognize an object. This time scale yields information about the algorithm used by the brain to detect objects. Studies that probe this time scale (speed-of-sight studies) performed with natural images are often limited because image content is determined by the photographer. These studies rarely contain systematic variations of scale, orientation and position of the target object within the image. Semi-realistic three-dimensional rendering of objects and scenes enables more systematic studies, allowing the isolation of specific parameters important for object recognition. To date, a computer vision algorithm that can distinguish between cats and dogs has yet to be developed and the specific cortical mechanisms that enable biological visual systems to make such distinctions are unknown. We perform a systematic speed-of-sight study as a step towards developing such an algorithm by enabling a better understanding of the corresponding biological processing strategies. In our study, participants are given the task of reporting whether or not a cat is present in an image ('cat / no cat' task). The object image is displayed briefly, followed by a mask image. As a mask, we use images of dogs as well as 1/f noise. We perform studies with natural images and with rendered images and compare the results.

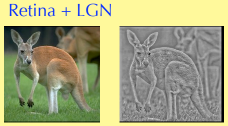
Objectives

- Constrain neural network model (feedforwad vs. feedback) by measuring time scale of human decision in 2 alternative forced-choice tests.
- Compare human performance for rendered images vs. natural images to enable generation of larger datasets
- Compare performance of human to performance of artificial neuro network

PANN: Peta-scale Artifical Neural Network

Object recognition is accomplished using a feedforward model of the primate visual system (Serre et al. 2007). Images are first fed through a model of Visual Area 1 (V1), the Visual Area 2 (V2), and finally classified using a model of the inferotemporal Cortex(IT). Both V1 and V2 feature simple (S1 and S2 respectively) cells, which respond to oriented bars and edges, and complex (C1 and C2 respectively) cells, which correspond to striate complex cells. Images are broken down into non-overlapping blocks of 16 pixels then passed to the S1 cells, which act as Gabor filters. Next, C1 cells group S1 cells that have the same preferred orientation. In this way, C1 cells become insensitive to the location and scale of the stimulus. At the S2 level, cells pool activities of retinotopically organized afferent C1 units with different orientations (feature tuning). The C2 layer groups S2 cells with similar orientations. The IT layer classifies images by taking the output of the C2 layer and running it through a Support Vector Machine. Using supervised learning, the SVM learns what features are found in a target image and searches for similar features in test images.

Retina + LGN



Local Contrast Adjustment

$$X_i \to \frac{X_i - \overline{X}}{\sigma}$$

$$\overline{X} = \frac{1}{N_{patch}} \sum_{i \in patch} X_i \qquad \sigma^2 = \frac{1}{N_{patch}} \sum_{i \in patch} X_i^2 - \overline{X}^2$$

V1 - Primary Visual Cortex

10^7 cells

100s features/receptive field

S cells : feature representation

C cells: local viewpoint invariance

V2 - prestriate cortex

V4

S cells : intermediate complexity features

C cells: increasingly larger viewpoint invariance

Greater Long-Term Plasticity

IT – Interotemporal Cortex

Classifier: e.g. Support Vector Machine (SVM)

Complex Object Identification

SET-UP

Examples of 3-D Rendering

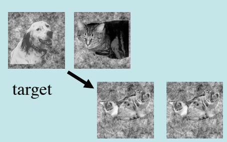
3-D rendering allows generation of massive customized dataset of target images and mask images

Examples of target and mask images

Experimental set-up

2 Alternative Forced Choice:

Subject must indicate which side the dog is on.

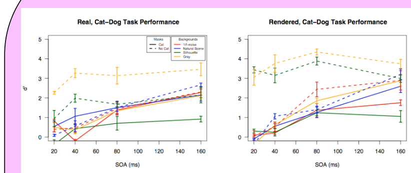


- 20, 80, 160 ms Stimulus Onset Asynchrony (SOA)
- 2 types of images -
  - segmented (background removed) real cats and dogs
  - Rendered (model) cats and dogs on transparent backgrounds
- 4 image modifications
  - Backgrounds added (gray, 1/f noise, natural images)
  - Images changed to a silhouette and placed on a gray background

Statistics of 5x5 pixel patches predicts number features/receptive field R_1*~100

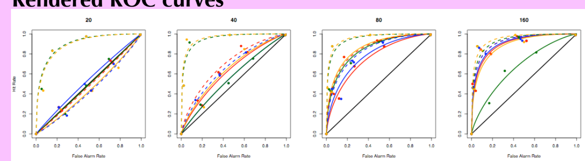
Hebbian learning builds **maximal entropy distribution** of size R_1*

RESULTS

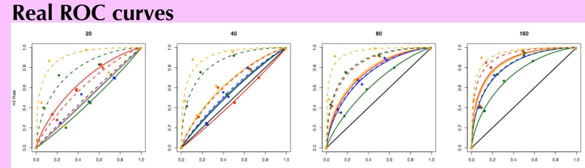


Left: Human psychophysics performance for four SOAs (20, 40, 80,1 60 ms). Masks were either the same as the background (dashed line), or same as background, with a cat overlaid on it (solid line). Below: Receiver Operating Characteristic (ROC) curves for real and rendered images at each SOA (n=5)

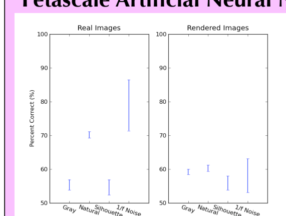
Rendered ROC curves



Real ROC curves



Petascale Artificial Neural Network (PANN)



Computer vision performance (10 trials, error bars denote standard error)

PANN was trained and tested using the same images that human psychophysics subjects were shown.

Best performance was with real cats and dogs on natural and 1/f noise backgrounds

Conclusions and Future Work

- Psychophysics tasks where the masks contained cats, were considerably more difficult for humans than tasks where the masks did not contain cats.
- Real images, on both natural and 1/f noise backgrounds, were easier for PANN to identify than rendered images on the same backgrounds
- PANN's performance on rendered images with different backgrounds was similar
- We are working to make PANN's performance more uniform between different backgrounds and real and rendered images.

This work was supported by DOE under project LDRD-DR- 2009006 and by the National Science Foundation.

Email:kys@lanl.gov, lmbettt@lanl.gov